High Performance Computing - 07

HDS-5230-07

Health Risk Prediction App

Done by:

Pravanith Reddy Kankanala

Supraja Medicherla

Venkata Sai Pallavi Pallapolu

Overview: The Health Risk Prediction App is a data science-based application that uses an XGBoost machine learning model to assess patients' health risks based on clinical and lifestyle variables such as age, gender, weight, calories intake, diabetic status and more. It helps patients and healthcare providers identify high-risk individuals and support early interventions.

Target Audience: Our application's primary target audience includes individuals with diabetes, heart disease, and those who engage in smoking or alcohol consumption. These groups are at elevated health risk and stand to benefit the most from personalized health monitoring and risk prediction. The model and features are tailored to identify and support these high-risk individuals to enable early intervention and promote healthier lifestyle choices.

Data source:

The dataset used to train this model was collected from:
https://www.kaggle.com/datasets/mahdimashayekhi/health-and-lifestyle-dataset

- **Description:** This dataset provides a well-structured collection of health and lifestyle-related data, covering various aspects such as physical activity, dietary habits, sleep patterns, mental well-being, and medical history.
- **Dataset Size:** 1000 rows and 16 columns
- **Features:** 'ID', 'Age', 'Gender', 'Height_cm', 'Weight_kg', 'BMI', 'Daily_Steps', 'Calories_Intake', 'Hours_of_Sleep', 'Heart_Rate', 'Blood_Pressure', 'Exercise_Hours_per_Week', 'Smoker', 'Alcohol_Consumption_per_Week', 'Diabetic', 'Heart_Disease'

Data Preprocessing:

Before training the prediction model, we prepared the health data using several cleaning and transformation steps to make it suitable for analysis.

- **Removed unnecessary columns**: We dropped the BMI column since we already had height and weight, which provide similar information. We also removed the ID column, which was just a unique identifier and not useful for prediction.
- **Categorical Encoding:** Gender was changed to numeric values: Female = 0, Male = 1. Smoker, Diabetic, Heart disease were also turned into numeric values: No = 0, Yes = 1.
- **Feature engineering:**

I. A new binary feature Hypertension was derived from the Blood_Pressure field. Individuals with systolic > 120 mmHg or diastolic > 80 mmHg were classified as hypertensive (Hypertension = 1), otherwise 0.

II. To enhance prediction accuracy, we engineered a new feature called Health_Risk_Score, which aggregates multiple lifestyle and physiological indicators into a single numeric risk score. This score helps reflect the overall health risk profile of an individual. We assigned points based on thresholds for various lifestyle and health variables:

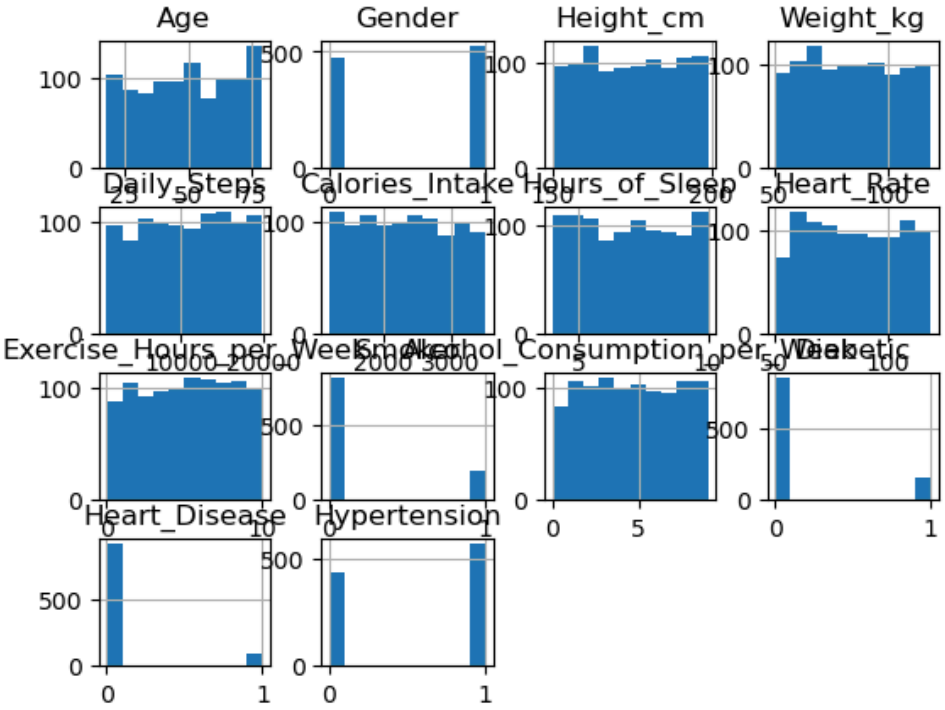
- Exercise per Week (hours): < 3 hours: +2 points (insufficient exercise), 3–5 hours: +1 point (moderate exercise), > 5 hours: 0 points (optimal exercise).[1]
- Daily Steps: < 5,000 steps: +2 points (low activity level), 5,000 – 9,999 steps: +1 point (moderate activity level), ≥ 10,000 steps: 0 points (ideal activity level).[2]
- Calories Intake: > 2,500 kcal/day: +2 points (high intake), 2,000 – 2,500 kcal/day: +1 point (moderate intake), < 2,000 kcal/day: 0 points (healthy intake).[3]
- Sleep Duration (hours/night): ≤ 7 hours or > 9 hours: +2 points (sleep out of ideal range), 7–9 hours: 0 points (ideal sleep duration).[4]
- Resting Heart Rate: < 60 bpm or > 85 bpm: +2 points (outside the healthy range), 60–85 bpm: 0 points (healthy range).[5]
- Alcohol Consumption per Week (Gender-specific):
  Female: > 7 drinks/week: +2 points (high alcohol intake), ≤ 7 drinks/week: 0 points (moderate intake)
  Male: > 14 drinks/week: +2 points (high alcohol intake), ≤ 14 drinks/week: 0 points (moderate intake).[6]

After calculating the score based on these factors, we classified individuals into three Health Risk Categories:

- 0 = Low Risk (score ≤ 2)
- 1 = Moderate Risk (score 3–5)
- 2 = High Risk (score ≥ 6)

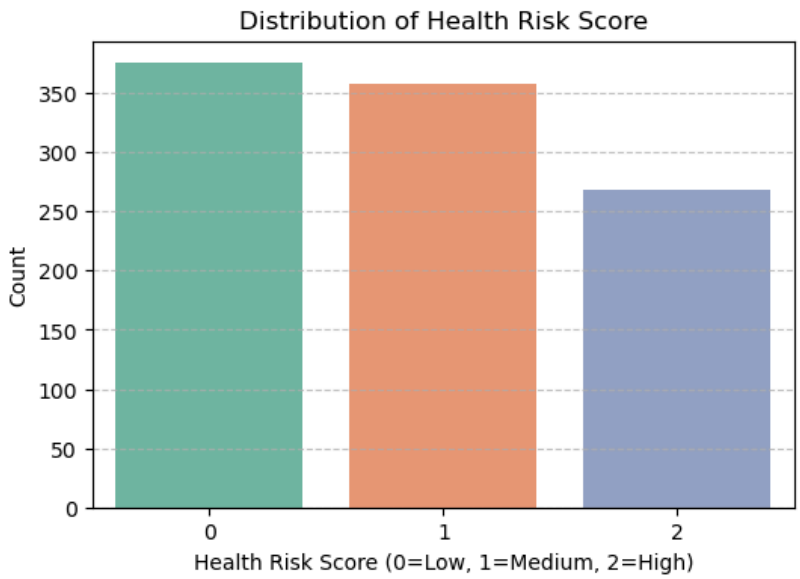
Exploratory Data Analysis (EDA):

- Histograms were plotted to observe the distributions of numerical and encoded features.



The histogram plots show that most numerical features (like Age, Height, Weight, and Calories Intake) are fairly well distributed, though a few have slight skewness. Categorical features such as Heart Disease, Hypertension, and Diabetic are imbalanced, with more individuals in the negative class. Gender appears balanced, and most individuals report low levels of alcohol consumption. These distributions help inform preprocessing decisions for model training.

- To understand the distribution of the Health Risk Score across the dataset, we plotted a count plot to show the frequency of each category (Low, Medium, High) in the Health Risk Score.



The distribution of health risk scores indicates that the dataset is relatively balanced across the three risk categories. The majority of individuals fall into the **low (0)** and **medium (1)** risk groups, with **high risk (2)** being the least represented. This balance supports robust model training without significant class imbalance.

Features and Target variable: After conducting Exploratory Data Analysis (EDA) and evaluating feature correlations, we selected Age, Gender, Height, Weight, Daily Steps, Calories Intake, Hours of Sleep, Heart Rate, Exercise Hours per Week, Smoker, Alcohol Consumption per Week, Diabetic, and Heart Disease as our predictor variables, and the Health Risk Score as the outcome (target) variable for our prediction model. The Health Risk Score is categorized into three classes: 0 – Low Risk, 1 – Moderate Risk, and 2 – High Risk. The predicted score determines the health risk category, which is displayed as Low, Moderate, or High.

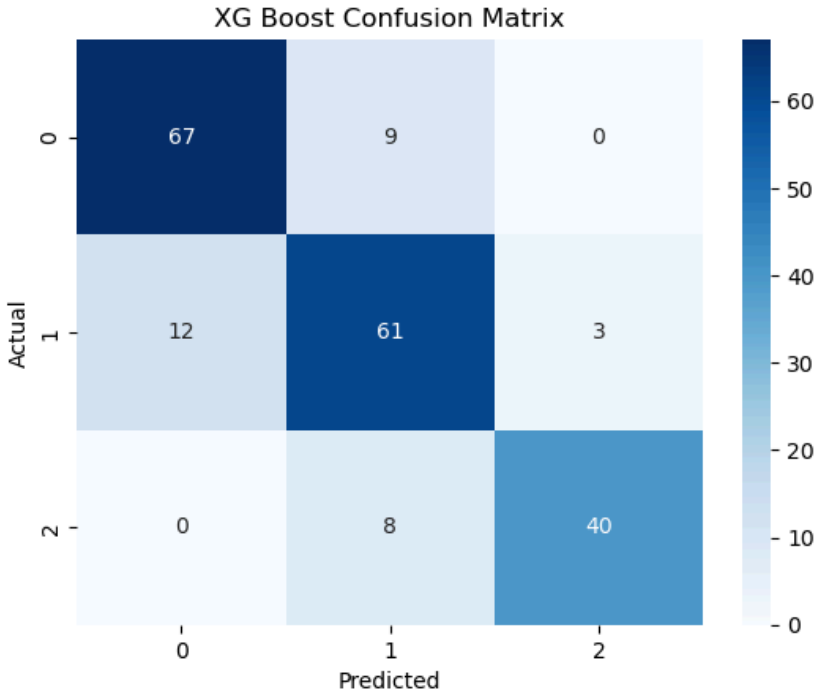
Model Implementation: XGBoost Classifier

For predicting the Health Risk Score, we used the XGBoost classifier on our dataset. Given that our goal is to predict health risk categories, which involve multi-class classification, XGBoost's high accuracy and ability to produce reliable predictions make it an ideal choice. The dataset was split into two subsets: 80% of the data was used for training, and 20% was reserved for testing. We ensured that the dataset was preprocessed correctly by excluding the target variable (Health_Risk_Score) from the feature set and confirmed that all remaining features were numeric.

Model Training: After initialization, the model was trained on the training data (X_train and y_train). The model learned patterns from the features to predict the Health Risk Score (which is categorized as Low, Medium, or High Risk).

Model Evaluation: After training the XGBoost classifier, we evaluated its performance on the test dataset using several standard classification metrics. The model achieved an accuracy of 84%, indicating that it correctly predicted the Health Risk Score category (Low, Moderate, or High) for the majority of the test cases. To assess the quality of predictions more comprehensively, we also calculated precision, recall, and F1 score using a weighted average to account for class imbalances. The model obtained a precision of 84.27%, a recall of 84%, and an F1 score of 84.05%, demonstrating a strong balance between sensitivity and specificity across all classes. These results suggest that the XGBoost model performs reliably in identifying different levels of health risk, making it a robust choice for multi-class health classification tasks.

Confusion Matrix: To further evaluate the model's performance, we plotted a confusion matrix using a heatmap. The confusion matrix provides a detailed view of the model's classification results by showing the number of correct and incorrect predictions for each class. Each row of the matrix represents the actual class (Low, Moderate, or High risk), while each column represents the predicted class.



The confusion matrix for the XGBoost model showed the performance of the classifier across the three Health Risk Score categories: 0 (Low Risk), 1 (Moderate Risk), and 2 (High Risk). The model correctly classified 67 out of 76 instances of Low Risk (class 0), 61 out of 76 instances of Moderate Risk (class 1), and 40 out of 48 instances of High Risk (class 2). A few misclassifications occurred; for example, 12 Moderate Risk instances were predicted as Low Risk, and 8 High Risk instances were predicted as Moderate Risk. Overall, the confusion matrix indicated strong model performance, especially in distinguishing between Low and High risk categories, though there was some overlap between adjacent risk levels particularly between Moderate and High Risk. This suggested that the model performed well.

App Development and Deployment: To create the application, we used Streamlit, an open-source Python library for building interactive web apps. The app was launched using the streamlit run command, which allowed us to quickly deploy our machine learning model and provide a user-friendly interface for inputting data and viewing health risk predictions in real time. To deploy the model, we first saved the trained XGBoost model as a .pkl file. We then created a project.py script that loads this model and handles predictions. A dedicated project folder was set up to organize the application files. Using the Anaconda Prompt, we installed Streamlit and launched the app by running the command streamlit run project.py. This allowed us to build an interactive web application for predicting health risk scores.

Citations:

1. American Heart Association. American Heart Association Recommendations for Physical Activity in Adults and Kids. American Heart Association. Published 2024. https://www.heart.org/en/healthy-living/fitness/fitness-basics/aha-recs-for-physical-activity-in-adults
2. Tudor-Locke C, Craig CL, Aoyagi Y, et al. How many steps/day are enough? For older adults and special populations. Int J Behav Nutr Phys Act. 2011;8:80. Published 2011 Jul 28. doi:10.1186/1479-5868-8-80
3. How Many Calories Should You Eat per Day to Lose Weight? Healthline. Published October 16, 2020. https://www.healthline.com/nutrition/how-many-calories-per-day#calorie-basics
4. Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. Prevalence of Healthy Sleep Duration among Adults — United States, 2014. *MMWR Morbidity and Mortality Weekly Report*. 2016;65(6):137-141. doi:https://doi.org/10.15585/mmwr.mm6506a1
5. American Heart Association. Target Heart Rates Chart. American Heart Association. Published March 9, 2021. https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates
6. Mayo Clinic. Alcohol in moderation: How many drinks is that? Mayo Clinic. Published December 11, 2021. https://www.mayoclinic.org/healthy-lifestyle/nutrition-and-healthy-eating/in-depth/alcohol/art-20044551